Exploring Mouse Tracking for Reading on Romanian Data

Maria Cristina Popescu, Sergiu Nisioi
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
popescu.cris.cristina@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

In this paper, we investigate the use of the Mouse Tracking for Reading (MoTR) method for a sample of Romanian texts. MoTR is a novel measurement tool that is meant to collect word-by-word reading times. In a typical MoTR trial, the text is blurred, except for a small area around the mouse pointer and the participants must move the mouse to reveal and read the text. In the current experiment, participants read such texts and afterwords answered comprehension questions, aiming to evaluate reading behavior and cognitive engagement. Mouse movement is recorded and analyzed to evaluate attention distribution across a sentence, providing insights into incremental language processing. Based on all the information gathered, the study confirms the feasibility of this method in a controlled setting and emphasizes MoTR's potential as an accessible and naturalistic approach for studying text comprehension.

1 Introduction

Language understanding is one of the most complex human cognitive activities. Whether reading or listening, the human brain processes linguistic input incrementally, integrating each word as it is encountered. This is known as incremental language processing and is characterized by both sequentiality and variability: some words are processed quickly, others require more cognitive effort due to low predictability, frequency, or syntactic complexity (Smith and Levy, 2013).

One of the main goals of psycholinguistics is to measure this incremental effort in real time. Early work in the 1970s introduced the *gaze-contingent moving window paradigm*, which involved making display changes in the text based on eye position as participants were reading, and then examining how these changes influenced eye movement behavior (McConkie and Rayner, 1975). Early studies demonstrated how parafoveal and foveal vision interact during reading and established the methodological foundation for modern incremental processing research. Eye-tracking provides the most precise measurement but is expensive and requires specialized equipment.

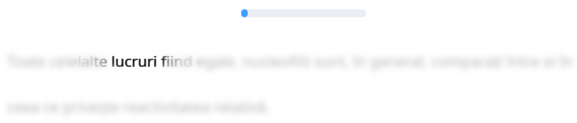
To address these limitations, alternative paradigms have been proposed. Self-paced reading (SPR) (Just et al., 1982) and the Maze task (Boyce et al., 2020; Forster et al., 2009) can be deployed online at low cost, but they involve linear reading and artificial constraints, affecting the interaction with the text.

In this paper we investigate the usage of Mouse Tracking for Reading (MoTR) method (Wilcox et al., 2024) on Romanian texts. The method was designed to balance the high accuracy and naturalness of eye-tracking with the low cost and online availability of other "self-paced" incremental measurements, such as SPR and Maze (Wilcox et al., 2024). In our MoTR experiment, participants are presented with several texts that are blurred, except for a small area around the mouse, which is clear. They have to move the mouse to reveal and read the text, and its position is recorded for post-processing (see Figure 1). After completing the reading of each sentence, the participants answer a comprehension question related to the text read, to validate the quality of the data and confirm the cognitive engagement of the candidate. Until now, published studies using MoTR have been conducted on English texts, often relying on corpora such as the Provo Corpus (Luke and Christianson, 2018; Wilcox et al., 2024), with relatively small documented applications in other languages (Schneider et al., 2021; Haveriku et al., 2025; Oğuz et al., 2025).

In this context, the present work addresses an important gap, being the first to test MoTR on Romanian texts.

The texts used in our experiment are Romanian

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025, pages 44–51, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-006



lucruri fiind

Figure 1: Example of the blurred interface used in the MoTR experiment. The only word visible in this image is “lucruri” (English *things*)

	Sentence len		Complexity	
	Eng.	Rom.	Eng.	Rom.
mean	22.82	24.46	0.24	0.13
std	7.74	8.62	0.19	0.25
min	7	9	0.02	0
max	45	49	0.93	1
samples	569		569	
sentences	158		158	

Table 1: Statistics comparing the English and Romanian Human Translation sentences (and 569 lexical complexity annotated samples). Romanian complexity annotations have a higher variance and a lower average complexity than the original English counterparts.

versions of the Multilingual Lexical Simplification Pipeline (MLSP) Shared Task 2024 competition dataset (Ghaddar et al., 2024a) The Romanian part has been created by manually translating the original English data such that each sentence contains similar lexical complexity annotations to the originals (Anghel et al., 2025). The data set is made so that it can be easily integrated into MultiLS (Ghaddar et al., 2024b), a recently developed framework for lexical analysis in multiple languages, providing a standardized context for comparative studies of text complexity and comprehension. Statistics regarding sentence length and annotated complexity scores are visible in Table 1.

An important advantage of this corpus is that certain words in each sentence are annotated with explicit human judgments of complexity scores assigned by five young adults. The complexity scores reflect the estimated difficulty of each word in its context. With this information, we analyze the relationship between the linguistic complexity of words and reading time, capturing the relation between perceived lexical difficulty and linguistic processing.

At the same time, we extend the existing experimental infrastructure by developing a complete pipeline in Romanian: from corpus preparation, to their integration into Magpie framework and the generation of comprehension questions. Overall, the contribution of the paper consists both in the methodological adaptation of MoTR for the Romanian language, and in demonstrating its applicability in the analysis of lexical complexity and reading times in an experimental setting.

2 Methodology

2.1 A MoTR Trial

In each trial of this experiment, participants are exposed to a web interface containing blurred text (see Figure 1), except for a small clear area around the tip of the mouse cursor. Each participant is instructed to move the mouse to reveal the text word by word, thus allowing sequential reading of the text. After the participant confirms the completion of reading a text by pressing a button, they are presented with a question with a “yes” or “no” answer, regarding the sentence read and intended to assess comprehension of the read content. At that moment, the entire text is blurred, no longer visible. Participants can move to the next screen only after answering the question.

Cursor movements are recorded throughout the reading, except for the moment when the participant answers the comprehension question. The cursor coordinates are subsequently analyzed as a proxy for gaze direction, effectively simulating the behavior of an eye-tracking system.

The experiment is implemented in Magpie[1], a web platform designed to conduct behavioral experiments directly in the browser. It allows for real-time transmission of cursor coordinates and task flow management.

2.2 Participants and Data Collection

Five native Romanian speakers (3F, 2M), 22-30 years old, agreed to participate in the experiment voluntarily. All participants are native Romanian speakers and have at least completed high school. None of them have diagnosed visual impairments. The study was conducted in a restricted and controlled environment, each session (approximately 2 hours per session) was directly monitored, to ensure that the rules and instructions were respected.

The data collected included:

- Mouse coordinates and timestamps

[1] Magpie is framework for building psychological online experiments that run in the participants’ browser: https://magpie-experiments.org/

- Word indices and reveal times
- Comprehension question responses
- Total reading duration per trial

To ensure that the MoTR method closely approximates the reader's visual attention, several parameters are calibrated:

- Spotlight size: 102 pixels - large enough to disambiguate word focus, but small enough to prevent excessive or fatiguing mouse movements;
- Gradual blur transition - simulating the shift from foveal to peripheral vision;
- Line spacing: 55px to avoid vertical interference;
- Cursor sampling rate: 20Hz -balancing temporal precision with transmission stability.

We make a small adaptation in terms of line spacing from the configurations proposed by Wilcox et al. (2024) so that users are less prone to accidentally move the mouse on the lines below or above the current reading areas.

3 Results

Our first objective is to provide an overview of how participants use the MoTR method and the variability that arises between individuals, items, and trials. At this stage, we focus on analyzing data from a single participant, selected due to their representative behavior. This case serves as an illustrative example of typical MoTR usage and provides a clear foundation for interpreting results in the broader analysis.

Total Reading Time (TRT) is used as our main measure of processing effort. It captures the full time spent on a word, including all refixations, and is widely used as an indicator of deep syntactic and semantic processing (Just and Carpenter, 1980; Rayner, 1998).

To ensure cognitive engagement and data quality, each sentence in the experiment is followed by a yes/no comprehension question. Participants show high accuracy, with individual scores ranging between 81% and 92%, and a group mean of approximately 88.5%.

This high level of accuracy confirms that participants have a high degree of comprehension, making the reading-time data more reliable. These results suggest that the MoTR interface supports natural reading and allows for meaningful variation in comprehension to be captured.

3.1 Correlation Analysis

We compute Pearson correlation coefficients to assess the linear relationship between total reading time and two basic lexical predictors:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \qquad (1)$$

where:

- x, y are numerical vectors of equal length n,
- m_x, m_y are the means of vectors x and y respectively.

Variable	Pearson *r*
Frequency Score	-0.53
Word Length	+0.58

Table 2: Pearson coefficients between *totalReadingTime* and lexical predictors. Frequency shows a moderate negative correlation with reading time, while word length shows a positive correlation. Frequency scores are obtained using the `wordfreq` library (Speer, 2022), which compiles word frequencies from diverse sources including Wikipedia, news, subtitles, and web data.

Pearson correlation analysis, as shown in Table 2, reveals a negative relationship between word frequency and reading time, and a positive one between word length and reading time. These findings align with well-established psycholinguistic assumptions: frequent words are processed more quickly, while longer words require more cognitive effort (Smith and Levy, 2013).

3.2 Regression Analysis

To estimate the influence of lexical and orthographic features on reading time, we use a regression model using *Support Vector Regression* (SVR), predicting continuous values by fitting a function within a margin of tolerance (Awad and Khanna, 2015). Although we employ a linear kernel, we opt for SVR instead of classic linear or Ridge regression due to its robustness in handling outliers and its ability to ignore small errors via the ϵ-insensitive loss function.

The model is implemented using the `SVR` module from the `scikit-learn` library (Pedregosa et al., 2011). The penalty parameter C is chosen via cross-validation and set to 100. Although this is a relatively large value, it consistently yields optimal predictive performance at cross-validation.

The SVR model predicts reading time as a linear combination of four features: frequency score, word length, syllable count, and the presence of diacritics. All features included in the model capture linguistic properties that influence processing difficulty. Frequency, word length, and syllable count are widely recognized as key factors influencing reading time (Rayner, 1998). We also include diacritics because omitting an accent can momentarily break the visual rhythm of a sentence and add a small cognitive load during speed reading (Marcet and Perea, 2022). Including diacritics in the model helps capture a subtle but systematic aspect of Romanian orthography that can influence reading behavior.

Each predictor is weighted by a learned coefficient β_i, and the model includes an intercept term β_0. Formally, the model takes the form: *totalReadingTime* $\approx \beta_1 \cdot$ FrequencyScore + $\beta_2 \cdot$ WordLength + $\beta_3 \cdot$ Syllables + $\beta_4 \cdot$ HasDiacritics + β_0.

Coefficient	Value
β_0	397.06
β_1	-34.39
β_2	186.38
β_3	-21.93
β_4	-9.58

Table 3: Estimated coefficients of the SVR model.

The intercept β_0 is the baseline reading time. β_1 shows that frequent words are read faster, while β_2 indicates that longer words take more time. β_3 and β_4 reflect smaller negative effects from syllable count and diacritics.

Model Performance

The SVR model is evaluated using 10-fold cross-validation, with the data split so that no sentences appears in both training and test sets. The model achieves a root mean square error (RMSE) of approximately **238.92 ms**. The coefficient of determination (R^2) is **0.37**, indicating that around 37% of the variance in reading times is explained by the model. The Pearson correlation between predicted and actual values is $r = 0.635$ ($p < 0.001$), suggesting a moderate and statistically significant fit.

Metric	Value
Coefficients	$[-34.39, 186.38,$ $-21.93, -9.58]$
Intercept	397.06
RMSE (mean, CV)	238.92 ms
R^2 Score	0.37
Pearson r	0.635 ($p < 0.001$)
Accuracy	87.24%

Table 4: Performance of the SVR model with a linear kernel and four predictors.

In addition to classic error metrics (RMSE, R^2), we evaluate model performance using the accuracy metric defined in (Hollenstein et al., 2022), where real and predicted values are scaled to [0, 100], and accuracy is defined as:

$$\text{Accuracy} = 100 - \text{MAE}$$

where MAE (Mean Absolute Error) represents the average absolute difference between predicted and actual values.

Our model achieves an accuracy score of **87.24%**, confirming a very good match between predicted and normalized real reading times.

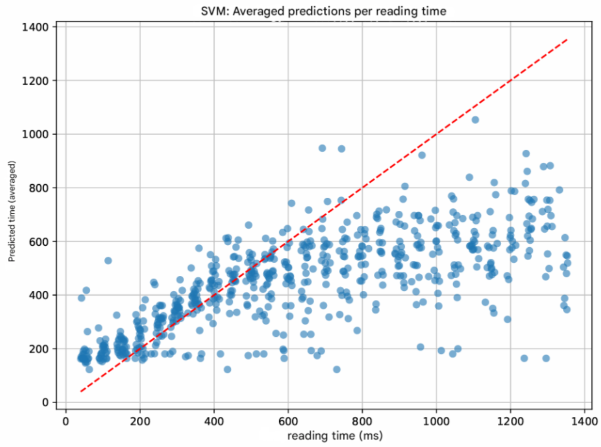


Figure 2: SVR predictions versus reading time. We observe a good alignment between predictions and actual values, with reasonable dispersion around the line $y = x$. We can observe a logarithmic tendency of reading times.

Figure 2 indicates a reasonable alignment between the predicted and actual values, with no obvious systematic deviations. The scatter around the identity line suggests natural variation in reading behavior.

The SVR model provides a flexible estimation of the relationship between word features and reading time, showing strong predictive performance in a cognitive-linguistic context (Li and Rudzicz, 2021).

3.3 Language Model Log-Probability

Linguistic surprisal is a computational measure of how unpredictable a word is in its context, and implicitly, how cognitively demanding it is to process (Smith and Levy, 2013; Hale, 2001). According to information theory, surprisal is defined as the negative logarithm of the conditional probability:

$$\text{Surprisal}(w_i) = -\log_2 P(w_i \mid w_1, w_2, ..., w_{i-1})$$

This equation reflects the idea that a highly expected word (high probability) requires less cognitive effort to process. In contrast, an unexpected word with low probability lead to higher surprisal values that typically requires longer reading time (Levy, 2008; Smith and Levy, 2013).

In masked language models such as BERT, surprisal is not based solely on the preceding context but instead uses the entire sentence. As such, we use the log-probability from the model as a proxy for the surprisal of a target word w_i.

Log-probability is calculated using the model *dumitrescustefan/bert-base-romanian-cased-v1*, a BERT-base model pre-trained on diverse Romanian data sources (Wikipedia, OSCAR, etc.) and adapted for masked language modeling tasks (Dumitrescu et al., 2020). The score is computed for each word in the sentences by masking it and retrieving the model's conditional probability. When a word is split into multiple subtokens during tokenization, we mask all subtokens simultaneously and compute the model's joint probability for the full word.

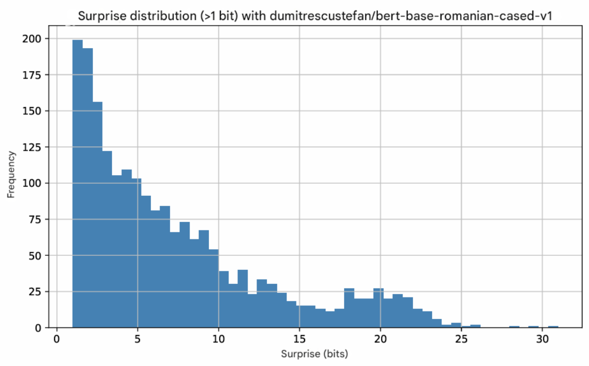


Figure 3: Distribution of log-probability values (> 1 bit) estimated using *bert-base-romanian-cased*. The distribution is right-skewed, with relatively few words showing high logprob.

For better visualization, we exclude very low surprisal values (below 1 bit), which dominated the frequency range and obscured the structure of more informative intervals.

To investigate the influence of surprisal on real-time processing, we analyze the relationship between estimated surprisal and total reading time per word (*totalReadingTime*). The computed Pearson correlation coefficients shows the following significant relationships:

Variable	**Pearson r**
Surprisal vs. Reading Time	+0.361
Frequency vs. Reading Time	−0.540
Word Length vs. Reading Time	+0.581

Table 5: Pearson correlation coefficients between predictors and reading time. Surprisal and word length correlate positively with reading time, while frequency correlates negatively.

These findings confirm that:

- Surprising words are associated with longer reading times;
- Frequent words are processed more quickly;
- Longer words tend to require more time to read.

To evaluate these predictors together, we fit a multiple linear regression model with surprisal, frequency, and word length as features. The estimated model predicts reading time as a linear combination of these three predictors. Each feature is multiplied by a learned coefficient (β_1, β_2, β_3), and the model includes a constant term β_0. Formally, the model takes the form:

ReadingTime $\approx \beta_1\cdot$ Surprisal + $\beta_2\cdot$ Frequency + $\beta_3\cdot$ WordLength + β_0.

The model is statistically significant ($F(3, 3890) = 853.5, p < 0.001$) and explains approximately 39.7% of the variance in reading times ($R^2 = 0.397$). The estimated coefficients are:

- $\beta_1 = +16.35$ (each additional bit of surprisal increases reading time by 16 ms),
- $\beta_2 = -45.71$ (higher word frequency reduces reading time),
- $\beta_3 = +57.30$ (each additional character increases reading time).

These results support the hypothesis that surprisal, frequency, and length contribute systematically to the cognitive effort involved in lexical processing (Levy, 2008).

3.4 Lexical Complexity and Reading Times

In addition to computationally derived predictors (surprisal, frequency, and word length), we also evaluate the relationship between a manually annotated measure of lexical complexity (*ht_complexity*) and total reading time. The *ht_complexity* values reflects human judgments of how difficult each word is to understand in its context, with higher values indicating greater perceived difficulty. The analysis is implemented by aligning annotated tokens with reading times from the dataset derived through manual translation and revision of the MLSP Shared Task 2024 corpus, as detailed in (Cristea and Nisioi, 2024).

The results indicate a significant positive correlation between lexical complexity and reading time (**r = 0.402**, $p < 0.0001$), suggesting that more complex units tend to require longer processing times.

This finding supports the hypothesis that lexical complexity directly impacts cognitive effort during reading, consistent with earlier work on linguistic processing and comprehension (Just and Carpenter, 1980).

3.5 Interindividual Variation in Reading Behavior

To evaluate the consistency of relationships between linguistic features and reading time, we extend our analysis to all five participants. This broader view provides more detailed insight into lexical effects and allows us to observe interindividual variability in language processing.

The average reading time (*totalReadingTime*) varies considerably across participants, with means ranging from approximately 435 to 604 milliseconds (Table 6).

In addition to the average reading times, the observed variability within each participant reflects clear differences in central tendency and dispersion. These results point to individual differences in reading styles and the stability of reading behavior (Just and Carpenter, 1980; Rayner, 1998).

We run separate regressions for each participant using word frequency and length as predictors. All show the same direction of effects—frequent words are read faster, longer words slower—despite variation in strength. This confirms that core lexical effects remain consistent across readers.

Participant	Mean	S.D.	Min	Max
P1	596.27	647.72	36.0	5751.0
P2	435.85	492.61	39.0	9011.0
P3	488.18	488.49	34.0	6349.0
P4	532.81	461.76	40.0	5400.0
P5	603.83	529.04	38.0	8005.0

Table 6: Descriptive statistics of reading times for each participant, including mean, standard deviation, minimum, and maximum values (all in milliseconds). Substantial differences can be observed across participants, both in mean and dispersion, suggesting variable reading styles.

4 BERT-based Predictor

We use the *bert-base-romanian-cased-v1* model (Dumitrescu et al., 2020), a pretrained version on large Romanian corpora that preserves the standard BERT architecture. The contextual embeddings generated by this encoder are integrated into a regression model, in order to predict the total reading time of a word based on the full sentence in which it appears.

Applying a logarithmic transformation (as suggested by the results in Figure 2) to the target value significantly improves model performance. This pre-processing step stabilizes the reading time distribution, reduces the influence of outliers, and allows the model to learn more robust relationships between contextual embeddings and cognitive reading difficulty.

We evaluate the final model on a test set of 773 examples, yielding the following metrics:

- Pearson correlation coefficient: **0.76**
- Spearman correlation coefficient: **0.78**
- Mean Absolute Error (MAE): **0.41** (in log space)
- Coefficient of determination R^2: **0.56**

These results confirm that large langauge models encode strong features for predicting reading times. The contextual embedding of the target token, combined with additional linguistic features and a log-transformed target, leads to accurate reading time predictions. Figure 4 shows a clear alignment between predicted and actual reading times, reflecting the model's strong predictive performance and robustness.

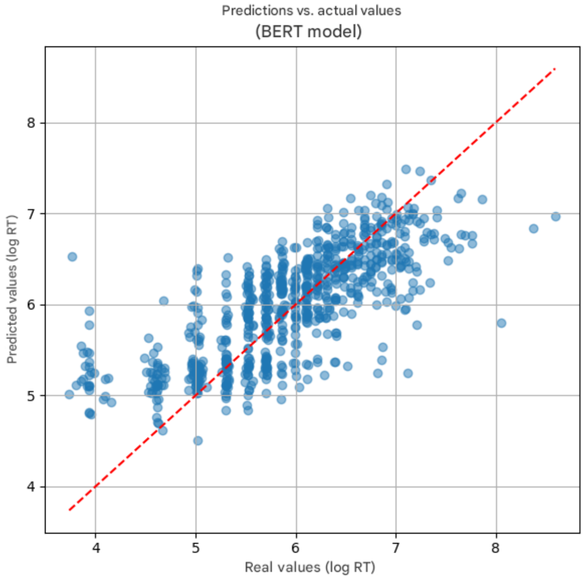


Figure 4: Predicted vs. actual reading times. The strong alignment along the diagonal suggests that the BERT-based model accurately predicts reading times from contextual embeddings.

5 Conclusions

This study shows that the *Mouse Tracking for Reading* (MoTR) method can be a practical and effective way to study how people read and process Romanian. Even though the number of participants was small, the results suggest that MoTR works well in controlled experiments.

One of MoTR's main advantages is its simplicity and accessibility. Because it runs in a web browser, it can be used both online and in physical locations, without the need of expensive equipment. While it doesn't offer the accuracy of eye-tracking, the blurred context outside the spotlight eliminates unwanted parafoveal effects, offering control over the text segments being read.

The statistical models confirm that reading times are strongly influenced by word length, frequency, and surprisal, findings that are in line with previous psycholinguistic research.

This research makes a new contribution by applying the MoTR paradigm in an experimental setting using Romanian, using a corpus adapted and validated for this task.

Future work involves expanding the experiment to a larger sample to increase confidence in results, a comparison between MoTR and traditional eye-tracking data, and the impact of time-guided lexical complexity predictions.

In conclusion, MoTR's ability to capture subtle aspects of cognitive processing during reading, along with its technical accessibility, makes it a strong alternative to traditional methods in experimental psycholinguistics.

Acknowledgments

This work was partially funded by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 (InstRead), and is supported by COST Action MultiplEYE, CA21131.

References

Fabian Anghel, Petru-Theodor Cristea, Claudiu Creanga, and Sergiu Nisioi. 2025. RALS: Resources and Baselines for Romanian Automatic Lexical Simplification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mariette Awad and Rahul Khanna. 2015. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA.

Veronica Boyce, Richard Futrell, and Roger P. Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Petru Cristea and Sergiu Nisioi. 2024. Archaeology at MLSP 2024: Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA2024)*, pages 610–617, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Kenneth I Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171.

Abbas Ghaddar et al. 2024a. MLSP 2024 shared task on multilingual linguistic and semantic proficiency. In *MLSP Shared Task*.

Abbas Ghaddar et al. 2024b. MultiLS: A framework for measuring linguistic and semantic complexity across languages. *arXiv preprint arXiv:2402.14972*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Alba Haveriku, Sara Bedulla, Nelda Kote, and Elinda Kajo Meçe. 2025. Understanding reading patterns of Albanian native readers through mouse tracking analysis. In *International Conference on Advanced Information Networking and Applications*, pages 433–443. Springer.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Marcel A Just, Patricia A Carpenter, and Jacqueline D Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of experimental psychology: General*, 111(2):228.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online. Association for Computational Linguistics.

Steven G. Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Ana Marcet and Manuel Perea. 2022. Does omitting the accent mark in a word affect sentence reading? evidence from spanish. *The Quarterly Journal of Experimental Psychology*, 75(1):148–155.

George W McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6):578–586.

Metehan Oğuz, Cui Ding, Ethan Gotlieb Wilcox, and Zuzanna Fuchs. 2025. Using MoTR to probe gender agreement in russian. *PsyArXiv*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

Cosima Schneider, Nadine Bade, Michael Franke, and Markus Janczyk. 2021. Presuppositions of determiners are immediately used to disambiguate utterance meaning: A mouse-tracking study on the german language. *Psychological research*, 85(3):1348–1366.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024. Mouse tracking for reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.